



US006287825B1

(12) **United States Patent**  
**Weissman et al.**

(10) **Patent No.:** **US 6,287,825 B1**  
(45) **Date of Patent:** **Sep. 11, 2001**

(54) **METHODS FOR REDUCING THE COMPLEXITY OF DNA SEQUENCES**

(75) Inventors: **Sherman Weissman**, New Haven; **Roger Lasken**, Guilford; **Xinghua Pan**, West Haven, all of CT (US)

(73) Assignee: **Molecular Staging Inc.**, New Haven, CT (US)

(\* ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/398,217**

(22) Filed: **Sep. 17, 1999**

**Related U.S. Application Data**

(60) Provisional application No. 60/100,999, filed on Sep. 18, 1998.

(51) **Int. Cl.**<sup>7</sup> ..... **C12P 19/34**; C12Q 1/68; G01N 33/00; C07H 21/02; C07H 21/04

(52) **U.S. Cl.** ..... **435/91.2**; 435/6; 435/91.1; 436/94; 536/23.1; 536/24.3; 536/24.33

(58) **Field of Search** ..... 435/6, 91.1, 91.2, 435/91.51, 183; 436/94; 536/23.1, 24.3, 24.33, 25.3

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

5,556,750 9/1996 Modrich et al. .  
6,010,850 \* 1/2000 Weissman et al. .... 435/6

**OTHER PUBLICATIONS**

Nelson et al., Genomic mismatch scanning: a new approach to genetic linkage mapping. *Nature Genetics* 4, Nov. 18, 1993.\*

Prashar et al., Analysis of differential gene expression by display of 3' end restriction fragment of cDNA. *Proc. Natl. Acad. Sci. USA* 93, 659-663, 1996.\*

Bloomquist et al., Rapid isolation of flanking genomic DNA using biotin-Rage, a variation of single-sided polymerase chain reaction. *DNA and Cell Biology* 11, 791-797, 1992.\*

*Biochemica* 3, 4-9, 1995.\*

New England Biolabs Catalog (1993/94), p. 21, Published by New England Biolabs, Inc. 32 Tozer Road, Beverly, MA 01915-5599, USA.\*

Nelson, *Nature Genetics* (1993), vol. 4: 11-18.

Straus et al. "Genomic subtraction for cloning DNA corresponding to deletion mutations" *Proc. Natl. acad. Sci.* vol. 87, pp. 1889-1893, Mar. 1990.

McAllister et al. "Enrichment for Loci Identical-by Descent between pairs of Mouse of Human genomes by Genomic Mismatch Scanning" *Genomics*, vol. 47, pp. 7-11, Jan. 1998.

Riley et al. "A novel, rapid method for the isolation of terminal sequences from YAC clones" *Nucleic Acid Research*, vol. 18, No. 10, pp. 2887-2890, 1990.

Cheung et al. "Genomic Mismatch Scanning: Applications to linkage and linkage disequilibrium analysis" *Am. J. of Human Genetics*, vol. 61, No. 4, Suppl. p. A271, Oct. 1997.

Prasher & Weissman *Proc. Nat. Acad. USA* (1996) 93: 659-663.

Geung et al. *Nature Genetics* (1998), 18: 225-230.

\* cited by examiner

*Primary Examiner*—Ethan Whisenant

*Assistant Examiner*—Frank LU

(74) *Attorney, Agent, or Firm*—Banner & Witcoff

(57) **ABSTRACT**

Genomic or cDNA, or fragments and mixtures thereof, can be screened by generation of subsets and then subjecting the subsets to mismatch scanning procedures. Alternatively, DNA fragments can be generated by cutting with a restriction endonuclease that generates variable overhangs. For either of the above methods, Y-shaped adapters having a region of non-complementary single-stranded DNA at the end can be used. Heterohybrid DNA, containing one DNA strand derived from each of two different samples, or homohybrids, containing DNA strands from the same sample, can be selected. Adapters attached to the ends of the fragments are designed to allow the selective isolation of homohybrid or heterohybrid DNA.

**25 Claims, 2 Drawing Sheets**

1 2 3



Fig. 1

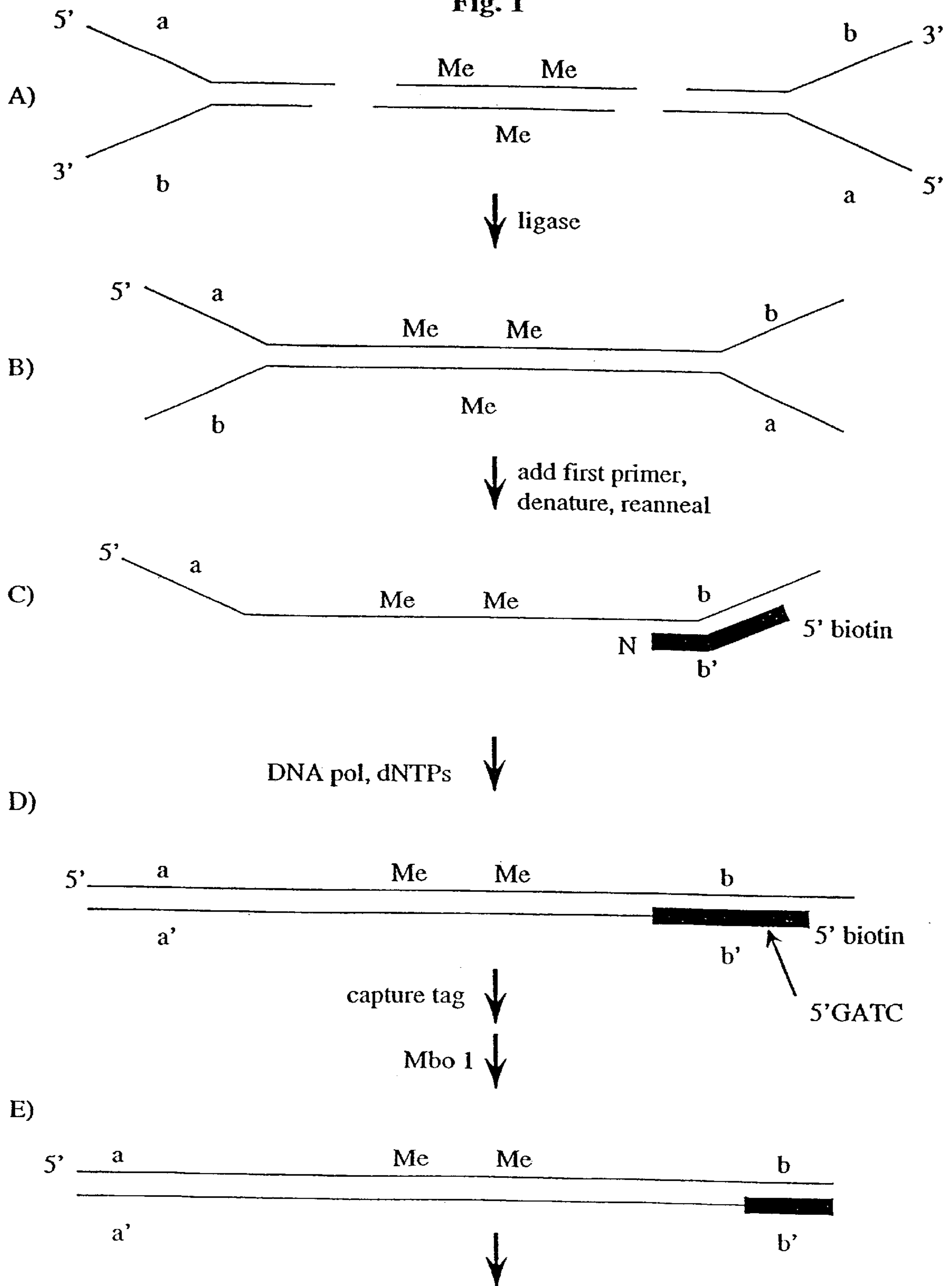
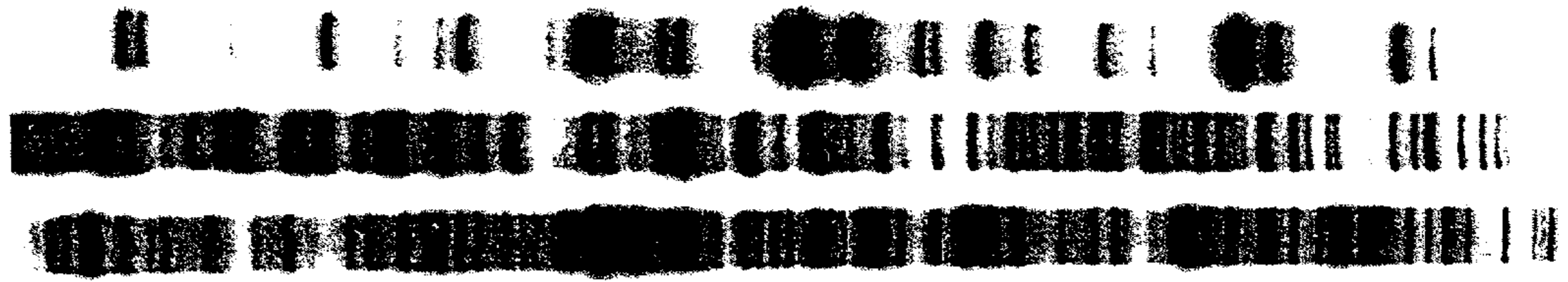


FIG. 2A

1 2 3

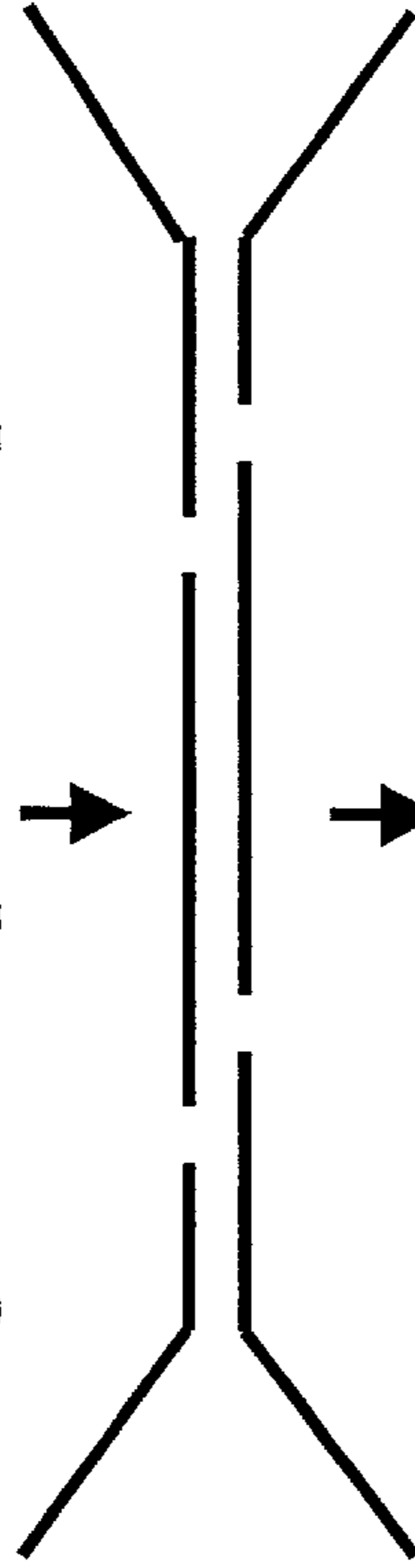


**Fig 2B cDNA Sau3A1 Fragments with Y shaped adapters. Amplification with Different Primer Pairs**

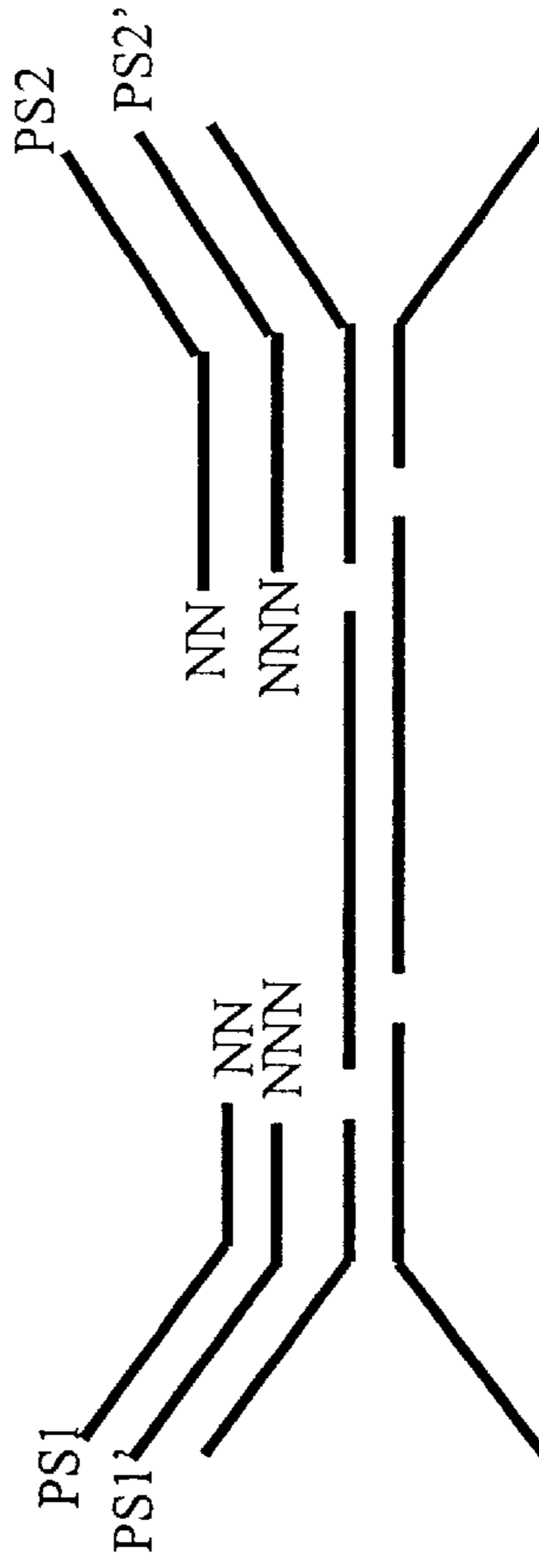
cDNA cut with Sau3A1



Ligate to a pair of adapters



PCR amplification with different primer pairs:



Well 1: PS1CG + PS2AG; Well 2:

PS1CGT+PS2AGG; Well 3: PS1CGT+PS2TCG

## METHODS FOR REDUCING THE COMPLEXITY OF DNA SEQUENCES

This application claims the benefit of application Ser. No. 60/100,999 filed Sep. 18, 1998.

### TECHNICAL FIELD OF THE INVENTION

This invention relates to methods for reducing the complexity of DNA mixtures, subsequent analysis of genetic variations, and isolation of probes or clones of regions of interest.

### BACKGROUND OF THE INVENTION

In 1993 Nelson and associates described a "genomic mismatch scanning" (GMS) method to directly identify identical-by-descent (IBD) sequences in yeast (Nelson, S. F., et al., *Nature Genetics*, 1993, 4:11-18; this and other papers, books and patents cited herein are expressly incorporated in their entirety by reference). The method allows DNA fragments from IBD regions between two relatives to be isolated based on their ability to form mismatch-free hybrid molecules. The method consists of digesting DNA fragments from two sources with a restriction endonuclease that produces protruding 3'-ends. The protruding 3'-ends provide some protection from exonuclease III (Exo III), which is used in later steps. The two sources are distinguished by methylating the DNA from only one source. Molecules from both sources are denatured and reannealed, resulting in the formation of four types of duplex molecules: homohybrids formed from strands derived from the same source and heterohybrids consisting of DNA strands from different sources. Heterohybrids can either be mismatch-free or contain base-pair mismatches, depending on the extent of identity of homologous regions.

Homohybrids are distinguished from heterohybrids by use of restriction endonucleases that cleave fully methylated or unmethylated GATC sites. Homohybrids are cleaved into smaller duplex molecules. Heterohybrids containing a mismatch are distinguished from mismatch-free molecules by use of the *E. coli* methyl-directed mismatch repair system. The combination of three proteins of the methyl-directed mismatch repair system MutS, MutL, and MutH (herein collectively called MutSLH) along with ATP introduce a single-strand nick on the unmethylated strand at GATC sites in duplexes that contain a mismatch (Welsh, et al., *J. Biol. Chem.*, 1987, 262:15624). Heterohybrids that do not contain a mismatch are not nicked. All molecules are then subjected to digestion by Exo III, which can initiate digestion at a nick, a blunt end, or a recessed 3'-end, to produce single-stranded gaps. Only mismatch-free heterohybrids are not subject to attack by Exo III; all other molecules have single-stranded gaps introduced by the enzyme. Molecules with single-stranded regions are removed by absorption to benzoylated naphthoylated DEAE cellulose. The remaining molecules consist of mismatch-free heterohybrids which may represent regions of IBD.

Nelson, et al., used *S. cerevisiae* hybrids as a model system and showed that sequences shared by two independently generated hybrids from the same parent strains could be identified in many instances. Experiments of this kind are much easier to do in yeast than in humans. The yeast genome is 250 times simpler than the human genome, it contains far fewer repetitive sequences, and genomic sequences of two yeast strains differ more than genomes of unrelated humans. It has thus far not been possible to do comparable experiments with human genomic DNA. In order to do so one

needs to use methods to reproducibly generate simplified but highly polymorphic representations of the human genome. Pooling techniques based on mathematical principles are also essential to identify IBD sequences as well as other sequences showing allele frequency differences (AFD) (Shaw, S. H., et al., *Genome Research*, Cold Spring Harbor Laboratory Press, 1998, 8:111-123).

The human genome is enormously long, at  $3 \times 10^9$  base pairs, and it is far too complex for efficient reannealing of homologous DNA strands after denaturation. The rate of annealing of a mixture of nucleic acid fragments in liquid phase is inversely proportional to the square of their complexity. Efforts have therefore been made to generate simplified representations of the genome for genetic methods based on cross hybridization of homologous sequences from different genomes. The exact degree of simplification of human genomic DNA needed to achieve efficient annealing depends on the conditions of hybridization including total DNA concentration, hybridization buffer, and temperature. In general a 10-100 fold simplification is needed for efficient annealing to occur at high DNA concentrations in high salt aqueous solutions (Lisitsyn, N. A., et al., *Science*, 1993, 259:946-951).

In some embodiments of the invention, DNA sequences of interest are replicated in rolling circle amplification reactions (RCA). RCA is an isothermal amplification reaction in which a DNA polymerase extends a primer on a circular template (Kornberg, A. and Baker, T. A., *DNA Replication*, W. H. Freeman, New York, 1991). The product consists of tandemly linked copies of the complementary sequence of the template. RCA can be used as a DNA amplification method (Fire, A. and Si-Qun Xu, *Proc. Natl. Acad. Sci. USA*, 1991, 92:4641-4645; Lui, D., et al. *J. Am. Chem. Soc.*, 1995, 118:1587-1594; Lizardi, P. M., et al., *Nature Genetics*, 1998, 19:225-232). RCA can also be used in a detection method using a probe called a "padlock probe" (Nilsson, M., et al., *Nature Genetics*, 1997, 16: 252-255).

It would be useful to have superior ways of analyzing human DNA and other complex DNA samples.

### SUMMARY OF THE INVENTION

A general method for screening genomic or cDNA, or fragments and mixtures thereof, involves sample simplification by the generation of subsets and then subjecting the subsets to mismatch scanning procedures. Any given DNA sequence will be represented in one and only one subset, minimizing the number of subsets required to detect a sequence of interest and guaranteeing that all possible sequences can potentially be covered by analyzing all possible subsets. The complexity of DNA sequences is reduced by attaching adapters to the ends of DNA fragments that allow the specific subsets of DNA to be selected and amplified. In some procedures, subsets are generated by replicating DNA in a polymerase chain reaction (PCR) or single primer extension reactions using primers that are complementary to sequences in the adapter and which, at the 3'-end, are complementary to a subset of sequences in the genomic or cDNA.

In another version of this method, DNA fragments are generated by cutting with a restriction endonuclease, such as Bsl1, that generates variable overhangs for which some of the nucleotides can have any of 2 to 4 of the bases A, C, G, or T. In this case, subsets are generated by ligating adapters to the fragment ends that have a specific sequence in the overhang and a primer binding site unique for each adapter. For either of the above methods, Y-shaped adapters can be

used having a region of non-complementary single-stranded DNA at the end. Therefore, following ligation, the DNA fragment-plus-adaptor construct has the non-complementary region at its ends. Use of Y-shaped adapters make it possible to generate non-overlapping subsets such that a given DNA

fragment will only be represented in one of the possible subsets. Procedures are given for isolating selected subsets from other, contaminating DNAs by using primers that have attached chemical moieties that can be captured on beads, columns, and the like. In some cases, the DNA is then released by cutting specifically designed sequences in the primers with restriction endonucleases. Fragment DNA is protected from these restriction endonucleases by methylation. The DNA subsets obtained are sufficiently reduced in complexity to allow improved analysis of sequence polymorphism by mismatch scanning procedures. Procedures are given for selecting DNA fragments representing regions of low polymorphism or for generating fragments depleted for regions of low polymorphism.

In some embodiments, the DNA fragments are replicated in a rolling circle amplification procedure (RCA; see reviews by Hingorani, M. M., and O'Donnell, M., *Current Biology*, 1998, 8:R83-86 and by Kelman, Z., et al., *Structure*, 1998, 6:121-5). The DNA polymerase III holoenzyme (hereafter sometimes denoted DNA pol III) is used in most of these methods to increase the rate and processivity of primer extension. DNA pol III also improves the ability to replicate through a DNA region of high GC content or other obstructions that tend to block DNA polymerases.

A method is also given for selecting heterohybrid DNA that contains one DNA strand derived from each of two different samples or homohybrids in which the DNA strands from different samples have not been recombined. Each DNA sample may consist of some concentration of a unique DNA fragment, or a mixture of fragments, and each sample may be derived from a single individual or more than one individual. The different DNA samples are mixed together, denatured, and then reannealed. Some of the DNA strands will reanneal back together with another strand from the same DNA sample forming a homohybrid. Other DNA strands will reanneal with a DNA strand from a different sample forming a heterohybrid. Adapters attached to the ends of the fragments are designed to allow the selective isolation of homohybrid or heterohybrid DNA. In one method, restriction endonuclease recognition sites are present in the adapters such that homohybrid or heterohybrid DNA can be selectively eliminated depending on the ability of the restriction endonuclease to cut the DNA.

#### BRIEF DESCRIPTION OF THE FIGURE

FIG. 1 is a diagram illustrating the addition of Y-shaped adapters to DNA fragments and generation of subsets reducing the complexity of the sequences. In FIG. 1A and FIG. 1B), Y-shaped adapters having non-complementary sequences on one end and appropriate overhangs for ligation on the other end are ligated to DNA fragments. In FIG. 1C, a primer is annealed to the denatured fragment-plus-adaptor construct for use in single primer extension, PCR or other DNA polymerase reaction. The 5'-end of the primer consists of a sequence complementary to the adapter region (b) and, at the 3'-end, the primer has one or more nucleotides (N) which must properly anneal to the fragment sequence in order for priming to occur. Therefore, only a subset of fragment sequences that are complementary to the nucleotide(s) N of the primer will be replicated. A capture

moiety, in this case biotin, can be present to allow isolation of reaction products. In FIG. 1D, extension of the primer by DNA polymerase generates a product, the 3'-end of which is complementary to the adapter region (a). Therefore, this DNA product can itself be replicated by use of a primer complementary to the sequence (a) in a primer extension, PCR, or other DNA polymerase reaction. Because of the Y-shaped adapters, the products of such replication reactions will be in non-overlapping subsets defined by the nucleotide (s) N of the primer. The presence of a restriction endonuclease recognition site in the adapter, in this case GATC (FIG. 1D), allows for the release of any DNA product following capture by the moiety as shown in FIG. 1E.

FIG. 2A shows cDNA Sau3A1 fragments with Y shaped adapters and amplification with different primer pairs.

FIG. 2B shows the results of amplification with three different primer pairs.

#### DETAILED DESCRIPTION OF THE INVENTION

This invention provides for the screening of complex DNA preparations, including complex DNA comprised of genomic segments or cDNAs, and the isolation of genes without requiring prior knowledge of their biochemical function or map position. Methods of the invention divide DNA into subsets and then manipulate the subsets using a mismatch repair system and capture techniques to obtain specific DNA sequences, including genomic subsets of long genomic DNA generated by selective amplification of sequences exhibiting low polymorphism.

As used herein, "polymorphism" refers to genetic sequence variation between different individuals of a species. A "homoduplex" is double-stranded DNA where both strands derive from the same genome or pools of genome samples, and a "heteroduplex" is double-stranded DNA where each strand originated from different genomes or different pools of genomes. By "perfectly matched" is meant double-stranded DNA where each base residue is correctly paired with a base on the opposite strand, i.e., A to T and C to G. By "mismatched" is meant double-stranded DNA where at least one base residue on either strand is either not paired with any residue, or paired with an incorrect base, i.e., A not paired with T, C not paired with G.

In a typical practice of a method of the invention, at least one DNA sample is methylated, usually at the GATC sites with bacterial DAM methylase, and the sample is then cut with an enzyme that makes infrequent cuts such as Pvu1 (Nelson, S. F., et al., cited above). Any type of DNA sample may be subjected to methods of the invention, including genomic DNA, genomic fragments, cDNA, cDNA fragments, and mixtures of any of these. It is an advantage of the invention that it can be used to identify identical-by-descent sequences of low polymorphism in complex human or other genomic DNA samples. It can also be used to identify sequences of high polymorphism.

Adapters are then ligated to the fragments to obtain fragment-plus-adaptor constructs. Linear or Y-shaped adapters may be employed. Y-shaped adapters are used in many preferred embodiments, but, in some cases, where Y-shaped adapters are illustrated, the methods can also be adapted to conventional linear adapters. Y-shaped adapters have been described (see Prashar, Y., and Weissman, S., *Proc. Natl. Acad. Sci. USA*, 1996, 93:659-663). A Y-shaped adapter typically has an overhang on its 3'-end for ligation, and on the 5'-end, a stretch of noncomplementary sequence on the opposite strands, giving rise to its Y-shape (see FIGS. 1A

and B). It is an advantage of the invention that, in preferred embodiments, the Y-shaped adapters allow for the synthesis of non-overlapping subsets of DNA. In typical embodiments, if the invention is carried out with conventional, linear primers, then the PCR-generated subsets will be partially overlapping, that is, some DNA sequences will be represented in more than one subset.

The fragment-plus adapter constructs are subjected to a PCR or to a single primer extension reaction in the presence of a primer complementary to at least a portion of the adapter at the 3'-end of the fragment-plus-adapter constructs and extending across the adapter ligation, and having at least one nucleotide overlap into the DNA fragment sequence. As used herein, a "polymerase chain reaction" includes conventional PCR, as well as modifications employing betaine, proof-editing polymerases, DMSO, and the like, and combinations thereof. Likewise, "rolling circle amplification" includes variants described by Hingorani and O'Donnell, cited above, and specifically encompasses modifications using a reconstituted bacterial polymerase III system including holoenzyme, helicase, clamp proteins, and clamp loading proteins (Bloom, L. B., et al., *J. Biol. Chem.*, 1997, 272:27919-27930).

In some cases, a tag at the 5'-end, and a restriction endonuclease recognition site at least about 6 nucleotides from the tag, are present to allow capture of a DNA product and subsequent release by cutting with the restriction endonuclease. In some embodiments, the annealed primer extends across the adapter ligation site one nucleotide into the DNA fragment sequence; in others, they extend two; and in others, more than two. The number of nucleotides, and the identity of the nucleotides that the primer extends across the adapter ligation site, determines the members of the subset to be amplified. The tag in many embodiments is biotin, illustrated in FIG. 1C.

In an alternative method for generating subsets, DNA samples are cut with a restriction endonuclease, such as BslI, that generates variable overhangs. That is, some of the bases in the recognition site can be of any two or more of the four possible bases G, A, T, or C. Adapters having overhangs complementary for this restriction endonuclease recognition site are ligated onto the fragments. Adapter overhangs having a unique sequence for the variable sites will only ligate to a subset of fragments that are complementary at those positions. Therefore, a subset of fragments will be replicated by a primer complementary to the adapter. By employing Y-shaped adapters, the subsets will be non-overlapping. Another advantage of this method is that it is a simple process to ligate adapters of one sequence at one end of the fragment and adapters of a second sequence at the other end of the fragment. If the adapters differ from each other in the primer annealing sequence of their non-complementary (Y-shaped) regions, then it is possible to amplify only one strand of the duplex adapter-fragment complex with the appropriate primer set in a PCR or other DNA polymerase reaction.

In a typical practice of a method of the invention, a subset of fragments are generated from one sample of a DNA or a mixture of DNAs, and these are methylated. The same subset is obtained from a second sample of DNA or mixture of DNAs, and these are not methylated. Mixing, denaturing and reannealing the methylated and unmethylated samples together generates hemimethylated heterohybrids, and, where a large number of DNA samples have been pooled together, most of the reannealed duplex DNA will be heterohybrids. The reannealing thus primarily results in perfectly matched heterohybrids or mismatched heterohybrids,

depending upon the degree of polymorphism of the samples. In some cases, the mismatched heterohybrids are then selected by binding of MutS to the mismatch or subjected to MutSLH, which nicks any that contain the mismatched base pairs expected for regions of high polymorphism (see U.S. Pat. No. 5,556,750 to Modrich, et al., Cheung, V. G., et al., *Nature Genetics*, 1998, 18:225-230, and the references cited therein).

In the case where samples are treated with MutSLK the nick that is generated in mismatched DNA is utilized to identify, isolate, amplify, or clone these fragments using a variety of methods that take advantage of the nick. In one case, a capture agent such as a biotin-tagged nucleotide is added onto the nick by terminal transferase or some other DNA polymerase and the nicked fragment is thereby isolated. Alternatively, the nicked strand can be removed by treatment with an exonuclease according to a published method (Nelson, S. F., et al., cited above). The surviving strand is then selected by DNA amplification or other methods. In another use of MutSLH nicked DNA, the 3'-OH of the nick serves as a primer for a DNA polymerase. Extension of the 3'-OH requires that the DNA polymerase utilize a duplex DNA template by a nick translation or strand displacement reaction. The newly synthesized DNA can be detected by the incorporation of a radioactively or fluorescently labelled nucleotide, or captured by the incorporation of a nucleotide appropriately tagged with a capture agent such as biotin. Also, extension of the nick where the Y-shaped adapter-fragment constructs of this invention are employed results in a DNA product which can be specifically replicated with unique primer sets in a PCR reaction or with a unique "splint oligonucleotide" in a rolling circle amplification. Referred to above, RCA is an isothermal amplification reaction in which a DNA polymerase extends a primer on a circular template (see Kornberg and Baker and other references cited above). The product consists of tandemly linked copies of the complementary sequence of the template.

In the case of RCA, the "splint oligonucleotide" is frequently a single-stranded sequence complementary to the ends of the DNA that results from extension of the nick such that denaturation of the DNA and annealing of the splint to the extended strand circularizes it. If the DNA is circularized such that its two ends are brought together at a nick, then the ends can be ligated together by DNA ligase forming a covalently closed circle. This DNA can then be amplified in an RCA. Another aspect of this invention is that DNA polymerase III holoenzyme derived from *E. coli* or other bacteria, including gram-positive and gram-negative bacteria, or related DNA polymerases from eukaryotes that have clamp (PCNA) and clamp loader (RFC) components (Kornberg and Baker, cited above) can be employed as the DNA polymerase in RCA. Use of DNA pol III is advantageous in many embodiments because pol III has a greater rate and processivity than other DNA polymerases and provides superior yield and ability to replicate long templates and templates having obstructions to DNA replication such as high GC content, or unfavorable secondary structure or sequence context. The *E. coli* dnaB and dnaC proteins or other helicases and the single-stranded DNA binding protein (SSB) can also be used to facilitate the reaction (Kornberg and Baker, cited above).

In another use of the nick generated in mismatched DNA by MutSLH, the mismatched DNA is discarded and the perfectly matched DNA can thereby be selectively amplified. For example, PCR primers, or a splint oligonucleotide in the case of RCA, can be used to amplify those DNAs not

nicked by MutSLH whereas nicked DNA cannot provide an intact DNA template.

The methods employed in this invention depend on the isolation of heterohybrid DNA in which the two strands are derived from two different DNA samples. This can be accomplished by published methods (Nelson, et al., cited above). Improved procedures that do not require methylation of fragment DNA are included in this invention. Sequences in the adapters are designed to allow selective cutting of homohybrid or heterohybrid DNA with restriction endonucleases. In some methods, the adapters contain two adjacent restriction enzyme recognition sites with specific methylation patterns such that heterohybrid and homohybrid DNAs can be distinguished by the ability of the methyl groups to block cutting by the restriction endonuclease. In other methods, partial restriction endonuclease recognition sequences are present in which the adapter contains mismatched bases. In this case heterohybrid and homohybrid DNAs can be distinguished by the elimination of the mismatches which allows restriction endonuclease to cut these sites.

## EXAMPLES

### Example 1

#### Procedure for creating cDNA or Genomic DNA Subsets

This example illustrates the use of PCR to amplify a subset of cDNA. The method can be used also for total genomic DNA or other mixtures of DNAs. The Y shaped adapters are designed to create "butterfly" ends on the construct (see FIG. 2). The Y type adapters enable only one strand but not the other strand of the fragment to be amplified. Also, both strands can be amplified separately. This is useful when the fragment contains a mismatched base pair and it is desirable to amplify the strands separately. The Y shaped adapters also enable the amplified duplex to be sequenced. The PCR primers are designed so that their 5' end is complementary to adapter sequence, but 1-3 nucleotides at their 3' end (designated by "N" in FIG. 2) must base pair with the target DNA insert. The target DNA sequences that will be amplified are determined by the identity of the 3' terminal nucleotides of the PCR primers. Therefore, only a subset of sequences will be amplified and the complexity of the sample will be reduced.

In this example, cDNA was cut with a 4 nucleotide-recognizing restriction enzyme, Sau3AI. The restriction enzyme was inactivated after digestion was completed by treating it at 65° C. for 30 minutes. The digested DNA was then purified by phenol chloroform extraction. Y shaped adapters were formed by annealing as follows: 1.3 nmol XS1, 1.3 nmol XS2, 5 mM Tris-HCl pH7.5 and 100 mM NaCl in 100 microliters volume at 94° C. for 10 min, cool down to 37° C. for 2 hrs, then 32° C. for 2 hrs, 30° C. for 2 Hrs, 28° C. for 2 Hrs, 25° C. for 2 Hrs and on ice. The fragments were then ligated to Y shaped adapters as follows: cut cDNA 0.1 micrograms, adaptor pair 0.2 micrograms/13 pmol, ligase 8u and 1x ligase buffer in 5 microliters at 16° C. overnight. A subset of the sequences was then amplified with a pair of PCR primers (see primer sequences below) in the following mixture: 2 microliters of 200-fold diluted ligated product from above, 2ul of 2 micromolar each primer, 0.75 units AmpliTag Gold DNA Polymerase (Perkin Elmer), 2 mM each dNTP and 1x DNA polymerase buffer supplied by the manufacturer. The PCR was done in a Perkin Elmer Cetus Gene Amp PCR System 9600 with the program:

95° C., 4 min  
five cycles  
94° C., 30 sec  
55° C., 30 sec  
72° C., 30 sec  
25 cycles  
94° C., 30 sec  
65° C., 30 sec  
72° C., 30 sec  
72° C., 5 min.

As indicated in FIG. 2, three different primer sets were used. Well number 1 contained a primer sequence which has C and G as the two 3' terminal nucleotides (designated PS1CG), and PS2AG which has A and G as the terminal nucleotides. Wells 2 and 3 of FIG. 2 contained primer sets with different terminal nucleotides as indicated. As expected, the three PCR reactions produced different band patterns determined by which primer set was used.

Adaptor for Sau3AI cutting

XS1 (22 nt)

CGTCCGGCGCAGCGACGGTCAG SEQ ID NO: 1

XS2 (29 nt)

GATCCTGACCGTCCATCTCTGTCGCAGCG SEQ ID NO: 2

PCR Primers (corresponding to above Sau3AI adaptors):

Set1:

PS1CG (28 nt)

CGTCCGGCGCAGCGACGGTCAGGATCCG SEQ ID NO: 3

PS2AG (31 nt)

CGCTGCGACAGAGATGGACGGTCAGGATCAG SEQ ID NO: 4

Set2:

PS1CGT (29nt)

CGTCCGGCGCAGCGACGGTCAGGATCCGT SEQ ID NO: 5

PS2AGGO (32 nt)

CGCTGCGACAGAGATGGACGGTCAGGATCAGG SEQ ID NO: 6

Set3:

PS1CGT (29 nt)

CGTCCGGCGCAGCGACGGTCAGGATCCGT SEQ ID NO: 7

PS2TCG (32 nt)

CGCTGCGACAGAGATGGACGGTCAGGATCTCG SEQ ID NO: 8

The above description is for the purpose of teaching the person of ordinary skill in the art how to practice the present invention, and it is not intended to detail all those obvious modifications and variations of it which will become apparent to the skilled worker upon reading the description. For example, there are numerous variations of steps in the overall procedures, and for preparing the probes. Variations in primers having larger overlap with DNA fragments and various amplification techniques, for example, have already been mentioned. Following selective isolation of duplex DNA, it could be transcribed with T7 or other appropriate RNA polymerase, and the RNA used as a direct probe, or reconverted into double-stranded DNA in some embodiments. It is intended, however, that all such obvious modifications and variations be included within the scope of the present invention, which is defined by the following claims. The claims are intended to cover the claimed components and steps in any sequence which is effective to meet the objectives there intended, unless the context specifically indicates the contrary.

The papers, books and patents cited herein are expressly incorporated in their entireties by reference.

---

 SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 8  
 <210> SEQ ID NO: 1  
 <211> LENGTH: 22  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: DNA  
 adaptor for cutting by Sau3A1 restriction endonuclease.  
 <400> SEQUENCE: 1  
 cgtccggcgc agcgacggtc ag 22

<210> SEQ ID NO: 2  
 <211> LENGTH: 29  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: DNA  
 adaptor for cutting by Sau3A1 restriction endonuclease.  
 <400> SEQUENCE: 2  
 gatcctgacc gtccatctct gtcgcagcg 29

<210> SEQ ID NO: 3  
 <211> LENGTH: 28  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding DNA adaptor shown in SEQ ID NO:1.  
 <400> SEQUENCE: 3  
 cgtccggcgc agcgacggtc aggatccg 28

<210> SEQ ID NO: 4  
 <211> LENGTH: 31  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding to DNA adaptor shown in SEQ ID NO:2.  
 <400> SEQUENCE: 4  
 cgctgcgaca gagatggacg gtcaggatca g 31

<210> SEQ ID NO: 5  
 <211> LENGTH: 29  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding to DNA adaptor shown in SEQ ID NO:1.  
 <400> SEQUENCE: 5  
 cgtccggcgc agcgacggtc aggatccgt 29

<210> SEQ ID NO: 6  
 <211> LENGTH: 32  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding to DNA adaptor shown in SEQ ID NO:2.  
 <400> SEQUENCE: 6



-continued

cgctgcgaca gagatggacg gtcaggatca gg

32

<210> SEQ ID NO: 7  
 <211> LENGTH: 29  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding to DNA adaptor shown in SEQ ID NO:1.

&lt;400&gt; SEQUENCE: 7

cgtccggcgc agcgacggtc aggatccgt

29

<210> SEQ ID NO: 8  
 <211> LENGTH: 32  
 <212> TYPE: DNA  
 <213> ORGANISM: Artificial Sequence  
 <220> FEATURE:  
 <223> OTHER INFORMATION: Description of Artificial Sequence: PCR primer  
 corresponding to DNA adaptor shown in SEQ ID NO:2.

&lt;400&gt; SEQUENCE: 8

cgctgcgaca gagatggacg gtcaggatct cg

32

What is claimed is:

1. A method for reducing the complexity of DNA sequences in a DNA sample by generating non-overlapping subsets of sequences which comprises:

- (a) digesting the DNA sample with a restriction enzyme to obtain DNA fragments;
- (b) ligating Y-shaped adapters to both ends of the fragments to obtain fragment-plus-adaptor constructs;
- (c) removing excess adapters that are not ligated to the fragments;
- (d) amplifying the fragment-plus-adaptor constructs by carrying out a PCR reaction with a first DNA primer that is complementary to one strand of the Y-shaped adapters and that, at the 3'-end of the primer, overlaps the DNA fragment by at least one base, and a second PCR primer that is complementary to at least a portion of the adapter region of the 3'-end of the DNA product formed by elongation of the first primer and which also anneals to the adapter region, and, at the 3'-end of the primer, overlaps the DNA fragment sequence by at least one base; and
- (e) identifying DNA products obtained in step (d) from each primer set as non-overlapping subsets, the primer sets differing by the 3'-nucleotides overlapping the fragment DNA, wherein each subset is less complex than the DNA sample.

2. A method according to claim 1 wherein the PCR reaction carried out in step (d) employs a primer that overlaps the DNA fragment by one base.

3. A method according to claim 1 wherein the PCR reaction carried out in step (d) employs a primer that overlaps the DNA fragment by two bases.

4. A method according to claim 1 which comprises the further step of carrying out at least one additional PCR reaction in the presence of primers that differ in the combination of 3'-terminal bases to obtain other subsets.

5. A method according to claim 1 which comprises the further step of carrying out at least one additional PCR reaction in the presence of primers that differ in the combination of one or more 3'-terminal bases to obtain other subsets.

6. A method for reducing the complexity of DNA sequences in a DNA sample by generating non-overlapping subsets of sequences which comprises:

- (a) digesting the DNA sample with a restriction enzyme to obtain DNA fragments;
- (b) ligating Y-shaped adapters to both ends of the fragments to obtain fragment-plus-adaptor constructs;
- (c) removing excess adapters that are not ligated to the fragments;
- (d) denaturing the fragment-plus-adaptor constructs and carrying out a single round of DNA replication using a DNA primer that is complementary to one strand of the Y-shaped adapters and that, at the 3'-end of the primer, overlaps the DNA fragment by at least one base;
- (e) denaturing the product of the reaction in step (d) and carrying out another single round of DNA replication using a second primer that has a tag attached to the 5'-end and that is complementary to at least a portion of the adapter region of the 3'-end of the DNA product formed by elongation of the first primer and which also anneals to the adapter region, and, at the 3'-end of the primer, overlaps the DNA fragment sequence by at least one base;
- (f) trapping the DNA product of (e) with a capture agent that binds the tag; and
- (g) identifying DNA products obtained in step (f) from each primer set as non-overlapping subset, the primer sets differing by the 3'-nucleotides overlapping the fragment DNA, wherein each subset is less complex than the DNA sample.

7. A method according to claim 6 wherein the tag is biotin and the capture agent is avidin or strepavidin.

8. A method according to claim 6 wherein prior to ligation the fragments are treated with a methylase that protects the DNA from being cut by a specific restriction endonuclease, and the primers further comprise the specific restriction enzyme recognition site which is not methylated so that the fragment can be removed from the tag following trapping by the capture agent by cutting with the corresponding restriction enzyme without damaging the fragment.

9. A method for reducing the complexity of DNA sequences in a DNA sample by generating non-overlapping subsets of sequences which comprises:

- (a) digesting the DNA sample with a restriction endonuclease that generates variable overhangs for which some of the sequence positions in the overhangs at cut sites can be any of the four deoxynucleotides;
- (b) carrying out separate Y-shaped adapter fragment ligations in which there are two different Y-shaped adapters in each ligation which differ from each other in the sequences of their non-complementary tails and in which the sequence of the adapter overhangs to be ligated to the fragments are unique for each ligation reaction;
- (c) removing the free adapters;
- (d) carrying out PCR reactions with the fragment-adapter constructs from each ligation using the corresponding primer sequences required to amplify one of the DNA strands of the fragment-adapter construct, wherein products of each PCR reaction are less complex than the DNA sample.

10. A method according to claim 9 wherein the endonuclease is BslI.

11. A method according to any of claims 1 to 10 wherein the steps are repeated with at least one other DNA sample.

12. A method according to claim 1 wherein the DNA subset derived from one sample is methylated, and the DNA subset derived from another sample is left unmethylated, and the hemimethylated products obtained from mixing, denaturing, and reannealing the methylated and unmethylated samples are incubated with a combination of proteins MutS, MutL, and MutH of the methyl-directed mismatch repair system (MutSLH) to identify mismatched subset fractions.

13. A method according to claim 12 wherein the DNA is methylated with bacterial DAM methylase, and the hemimethylated products are cut with Dpn1 and Mbo1 prior to incubation with MutSLH.

14. A method according to claim 11 wherein the DNA subsets derived from the samples are mixed, denatured, reannealed, and incubated with *E. coli* MutS protein to identify mismatched subset fractions.

15. A method according to claim 14 wherein the mismatched fractions are further treated by at least one additional incubation with MutS protein to obtain at least one enriched mismatched subset fraction.

16. A method for carrying out claim 12 in which the hemimethylated DNA duplexes obtained are treated by methylating one sample prior to subjecting it to the method of claim 9, or ligations with linear adapters followed by denaturing, mixing the samples, and identifying heterohybrid hemimethylated DNA duplexes obtained and either:

- (i) (a) blocking the duplex ends with a modified nucleotide resistant to exonuclease activity;
- (b) nicking the duplexes obtained in step (b) with MutSLH;
- (c) digesting the nicked duplexes with an exonuclease; and
- (d) isolating the single-stranded DNA generated by the exonuclease; or
- (ii) (e) blocking the 3' ends of the fragments with terminal transferase and dideoxynucleoside triphosphates;
- (f) treating the blocked fragments with MutSLH to obtain nicked fragments;
- (g) attaching a tag to the nicks with a terminal transferase;

(h) trapping the tagged fragments with a capture agent to obtain subsets;

(i) melting off the complementary strands; and then

(iii) (j) amplifying the single stranded DNA obtained in either (i) or (ii) with a polymerase chain reaction or replicating the DNA in a single primer extension reaction; or

(k) employing adapters containing a specific sequence of an RNA polymerase promoter and selectively transcribing the remaining duplex DNA with the corresponding RNA polymerase and directly probing with the RNA or reconverting back into double-stranded cDNA; or

denaturing the duplexes, diluting them, annealing them to an excess of a splint oligonucleotide containing sequences complementary to both arms of the Y-adapters or conventional adapters attached to the 3'-end of the unmethylated DNA fragments, ligating the single-stranded DNA sufficiently diluted such that the ligation of the ends of a DNA strand will occur only when they are held together by the splint oligonucleotide, and then amplifying the resulting circular DNA using the rolling circle method, using the splint oligonucleotide or an oligonucleotide complementary to an adapter which has been elongated with a DNA polymerase capable of strand displacement.

17. A method according to claim 16 wherein single-stranded DNA is either partially or completely removed after digesting with an exonuclease.

18. A method according to claim 17 wherein the step is carried out with a benzoylated naphthoylated DEAE-cellulose column.

19. A method according to claim 16 wherein products obtained in steps (i) or (ii) are amplified in step (iii) by transcription using a T7 RNA polymerase promoter as an adapter.

20. A method for selectively isolating mismatched or perfectly matched DNA from the subsets obtained in claim 16 comprising either:

(i) (a) isolating the single-stranded DNA generated from mismatched DNA by the exonuclease; and

(b) amplifying the single stranded DNA with a polymerase chain reaction using one primer complementary to the 3'-hydroxyl-ended arm of the methylated DNA strand and another primer containing sequences from the 5'-phosphate terminated arm of the methylated DNA strand or replicating the DNA in a single primer extension reaction; or

(ii) (c) attaching a tag to the nicks with terminal transferase;

(d) trapping the tagged fragments with a capture agent to obtain subsets;

(e) obtaining complementary strands by melting them off; and then either

(f) carrying out step (b); or

(iii) (g) extending the nick that MutSLH created in mismatched DNA duplexes using a DNA polymerase;

(h) amplifying the extended fragment by PCR using one primer complementary to the appropriate arm of the extended fragment and another primer containing a sequence from the appropriate arm of the Y-adapter on the extended fragment; or

## 15

- (i) capturing perfectly matched DNA by denaturing the duplexes, diluting them and then annealing them to an excess of a splint oligonucleotide containing sequences complementary to one arm of the Y-adapter attached to the 3'-end of the unmethylated DNA fragments, and adjacent sequences complementary to the arm of the Y-adapter that had been attached to the 5'-end of the unmethylated DNA strand in heteroduplexes; and
- (j) ligating the single-stranded DNA diluted sufficiently such that the ligation of the ends of a DNA strand will occur only when they are held together by the splint oligonucleotide; and then
- (k) using either the splint oligonucleotide or a second oligonucleotide complementary to the single-stranded circle to amplify the circularized DNA by the rolling circle method.

## 16

21. A method according to claim 20 wherein the tag employed in (ii) is biotin and the capture agent is avidin or strepavidin.

22. A method according to claim 20 wherein the rolling circle amplification in (iii) is accomplished using a reconstituted bacterial polymerase III holoenzyme system that includes clamp proteins, and clamp loading proteins.

23. A method according to claim 22 wherein the holoenzyme system further comprises a helicase.

24. A method according to claim 16 wherein amplification using a polymerase chain reaction is employed.

25. A method according to claim 16 wherein the tag employed in (ii) is biotin and the capture agent is avidin or strepavidin.

\* \* \* \* \*

UNITED STATES PATENT AND TRADEMARK OFFICE  
**CERTIFICATE OF CORRECTION**

PATENT NO. : 6,287,825 B1  
DATED : September 11, 2001  
INVENTOR(S) : Sherman Weissman et al.

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Title page,

Item [56], **References Cited**, OTHER PUBLICATIONS, "Nov. 18" has been replaced with -- 11-18 --; "biotin-Rage" has been replaced with -- biotin-RAGE --; "Genetrics" has been replaced with -- Genetics --; and "of" in between "Mouse" and "Human" has been replaced with -- or --.

Column 12,

Line 54, "subset" has been replaced with -- subsets --.

Column 13,

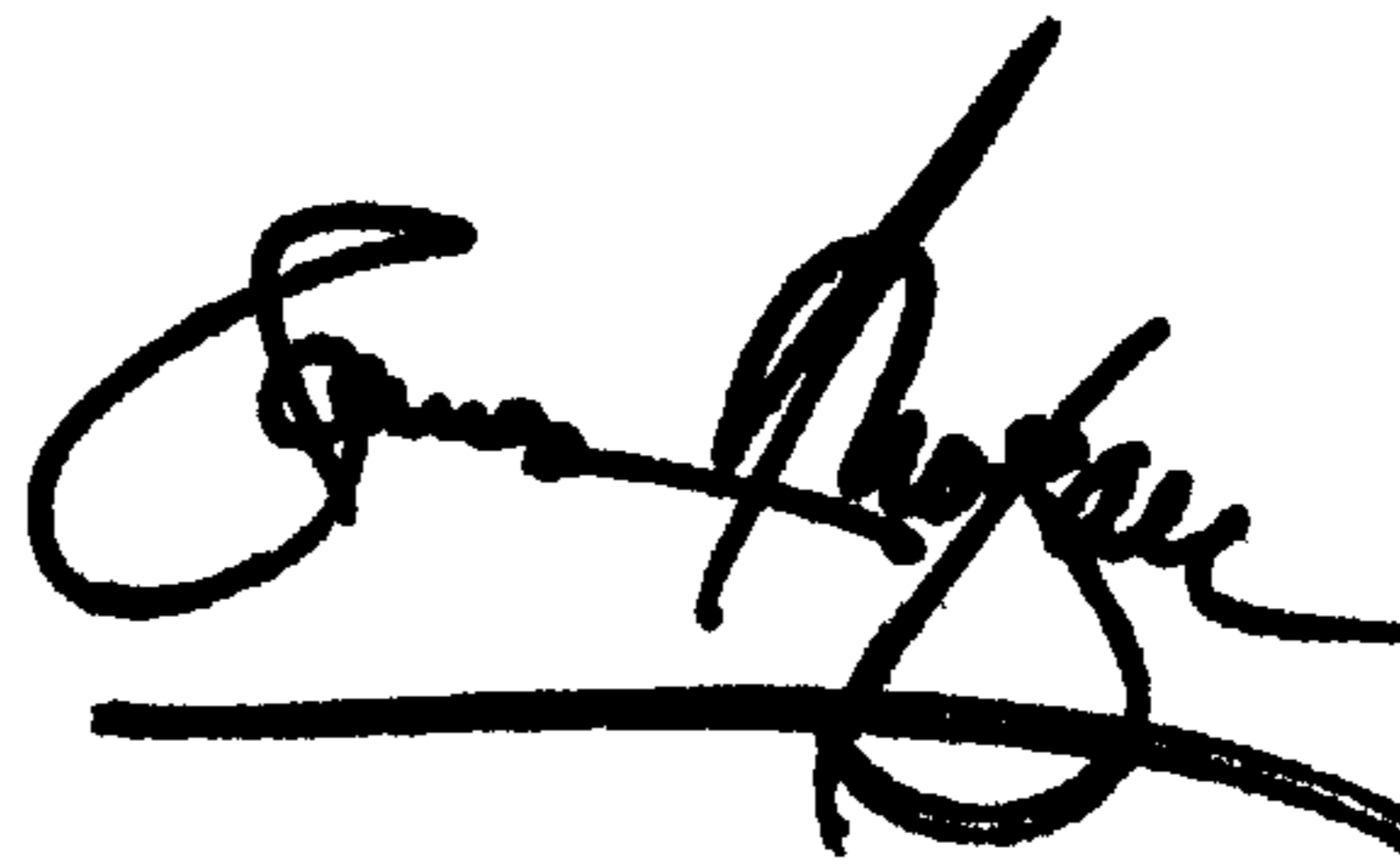
Line 23, "BslI" has been replaced with -- *BsII* --.

Line 38, "Dpn1" has been replaced with -- *Dpn1* --, and "Mbo1" has been replaced with -- *Mbo1* --.

Signed and Sealed this

First Day of October, 2002

*Attest:*



*Attesting Officer*

JAMES E. ROGAN  
*Director of the United States Patent and Trademark Office*

UNITED STATES PATENT AND TRADEMARK OFFICE  
**CERTIFICATE OF CORRECTION**

PATENT NO. : 6,287,825 B1  
DATED : September 11, 2001  
INVENTOR(S) : Sherman Weissman et al.

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

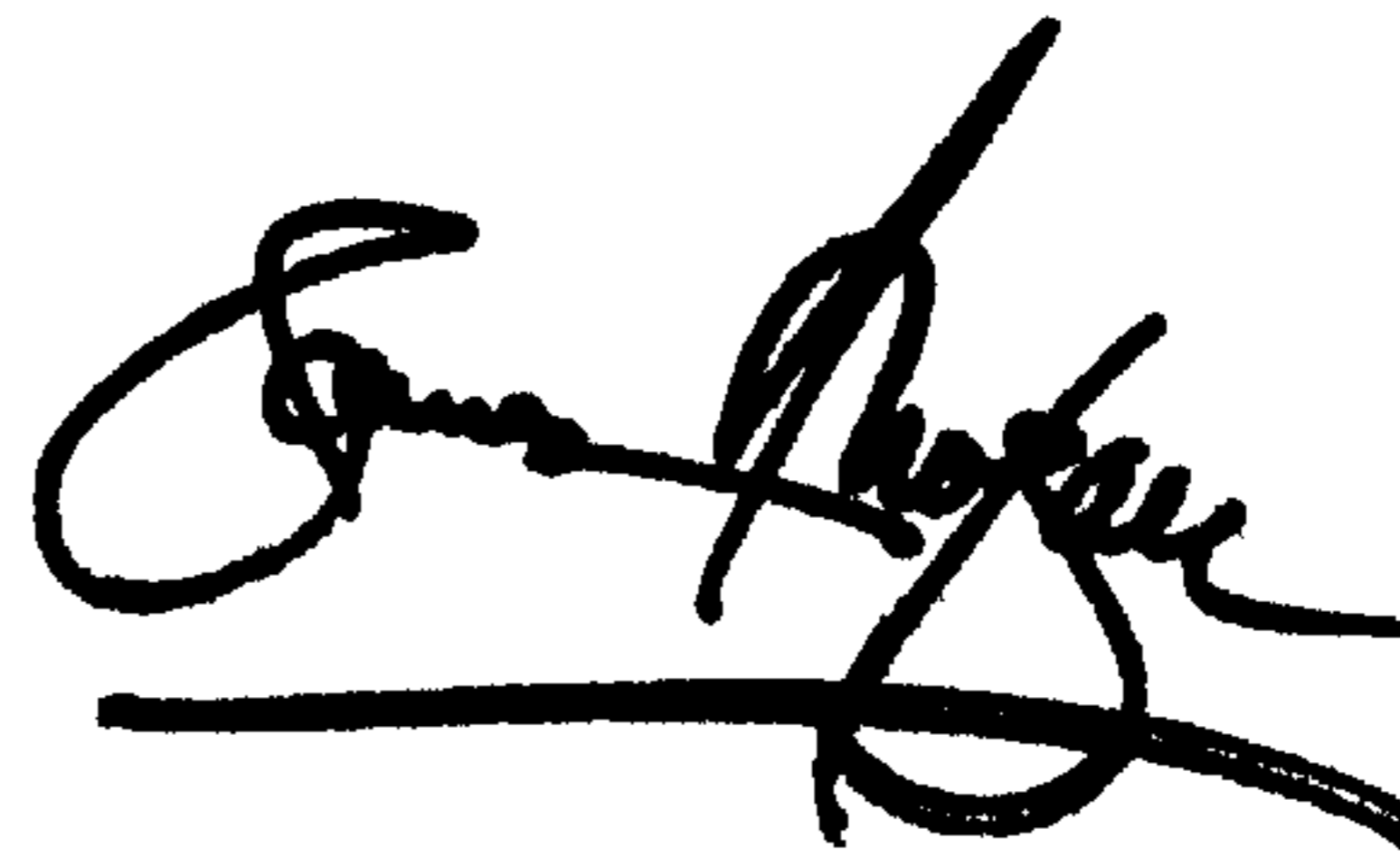
Title page,

Item [73], Assignee: "New Haven" has been replaced with -- Guilford -- and -- **Yale University**, New Haven, CT (US) -- has been added.

Signed and Sealed this

Twelfth Day of November, 2002

*Attest:*

A handwritten signature in black ink, appearing to read "James E. Rogan", written over a horizontal line.

*Attesting Officer*

JAMES E. ROGAN  
*Director of the United States Patent and Trademark Office*